Perceptual Evaluation of Dysphonic Voices: Can a Training Protocol Lead to the Development of Perceptual Categories?

*Alain Ghio, *Sophie Dufour, †Aude Wengler, *Gilles Pouchoulin, *,†Joana Revis, and *,†Antoine Giovanni,
*Aix-en-Provence and †Marseille, France

Summary: The purpose of this study was to develop and test a training protocol for the perceptual evaluation of dysphonia. A group of 38 inexperienced listeners participated in a three-phase experiment: a pretest to evaluate their initial performance on categorization of dysphonic voices, a training phase, and a posttest to detect training-related changes in performance. In parallel, a different group of 14 listeners who were experts in voice assessment took a test that was identical to the posttest taken by the inexperienced subjects. The corpus used for the tests was made up of recordings of 142 voices of women reading aloud, with a sampling of voice qualities ranging from normal to severely degraded. The learners' performance on judgments of moderate and severe dysphonia improved between the pretest and the posttest. No improvement was observed for normal voices, whose initial detection was already good, nor for slight dysphonias, which appear to be the most difficult to learn. The improvements were still present on a delayed posttest taken a week later. Unexpectedly, the inexperienced listeners' initial performance was similar to that of the experts. After the training phase, their scores for severely deteriorated voices were even better than the experts'. In conclusion, our training protocol seems to be effective and could therefore be proposed to voice therapists. However, judging intermediate degrees of dysphonia remains fragile and therefore needs to be reinforced by repeated training.
Key Words: Perceptual evaluation of dysphonia–Training–Voice–Expertise.

INTRODUCTION

Perceptual evaluation of dysphonia

To treat patients with dysphonia, it is essential to assess the quality of their voices. Such patients usually decide to consult a voice specialist when they begin to hear changes in their own vocal output. Similarly, after having undergone surgical treatment or speech therapy, they generally judge the treatment's success in terms of the auditory impression they have of their voice.[1] Perceptual evaluation is the most widespread method used by clinicians to describe a patient's voice (breathy, hoarse, rough, etc.) or measure the severity of the dysfunction. This method was recommended by Dejonckere et al[2] in their basic protocol for the functional evaluation of voice pathologies. It has many advantages: it is easy to implement, inexpensive, and directly accessible to any clinician. However, although perceptual analysis remains the standard in this field, it nevertheless raises a key question: Just how reliable is it? Reports in the literature indicate substantial variability in perceptual judgments of the voice.[3–9] Variability shows up as inconsistencies between ratings of the same voice made by different listeners (between-listener variability) and between ratings made by the same listener at different times (within-listener variability).

Assessment variability has been widely studied in view of alleviating these phenomena, deemed undesirable from the clinical standpoint. Various authors have taken different approaches in an attempt to reduce its magnitude, including recruitment of experts vs naive listeners,[6] the use of analog vs categorical scales,[7] judgments of different types of utterances (sustained vowels vs sentences),[8] and ratings along various dimensions (overall quality vs breathiness[9] vs roughness).

Variability vs reliability

As a general rule, studies in this area have focused on observing variability-related phenomena, considered indicative of the assessment method's lack of reliability.[10] We think it is preferable to test method reliability directly by measuring the accuracy of the listeners' responses. Although it seems legitimate to assume that reliable responses exhibit little variability—because, by definition, a correct response is invariable—the reverse is not trivial. For example, if dysphonia severity is being rated on a four-level scale like Hirano's,[11] where G0 is a normal voice, G1 a slightly dysphonic voice, G2 a moderately dysphonic voice, and G3 a severely dysphonic voice, then listeners who systematically give a rating of G0 for globally normal voices and G2 for globally abnormal voices would obtain a low degree of variability because of their simplified response strategy. The responses might also be highly valid, if these listeners are correctly rating normal voices as G0 and dysphonic voices as G2. But their responses could not be considered as sensitive enough because they did not use all four levels of the rating scale.

According to Bele,[5] the reliability of an evaluation is related to the degree to which the results are void of measurement errors. This author makes the distinction between random errors (listener distraction, poor use of the response choices, etc.) and systematic errors. Although the former type of error can be minimized by repeating the tests and increasing the number of participants, the second affects the listener's score and

Accepted for publication July 14, 2014.
From the *Laboratoire Parole et Langage, Aix-Marseille University, CNRS, Aix-en-Provence, France; and the †Department of Otolaryngology — Head and Neck Surgery, Timone University Hospital, Marseille, France.
Address correspondence and reprint requests to Alain Ghio, Laboratoire Parole et Langage, Aix-Marseille University, CNRS, 5 Avenue Pasteur, BP 80975, 13604 Aix-en-Provence Cedex 1, France. E-mail: alain.ghio@lpl-aix.fr
Journal of Voice, Vol. 29, No. 3, pp. 304-311
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.07.006

reveals an idiosyncrasy of that particular participant. If a systematic error is made by all listeners, then this points either to a specific feature of the stimulus or to a limitation of the evaluation method.

The key issue raised in the present study, then, concerns how to categorize dysphonic voices. What is a normal voice? What is a slightly, moderately, or severely dysphonic voice? This question is all the more complex because of the multiple manifestations of dysphonia—a dysphonic voice can be breathy, hoarse, or rough, diplophonic, hyperfunctional, or hypofunctional, and so on—making it difficult to define clear-cut categories, whether in terms of quality or quantity. This brings us to the topic of category exemplars and prototypes.

Categorization, exemplars, and prototypes

Dysphonia evaluation involves a categorization process (assigning a grade to a vocal production), which requires:

- being able to put similar vocal productions into the same category and
- being able to distinguish between vocal productions that belong to two different categories.

According to the so-called "classical" theory, a category is defined by the properties its members share, and any entity that possesses those properties belongs to the said category: if an exemplar possesses a given property then it belongs to category X; if it does not possess that property, then it does not belong to category X. This approach was questioned by Rosch,[12] whose experiments showed that in most cases, it is difficult to define a set of necessary and sufficient properties to characterize a category. The most common example is the ostrich, which has many of the defining features of the bird category (two feet, two wings, feathers, a beak, lays eggs, etc.), but does not fly. Having noted also that certain exemplars are more representative of a category than others, Rosch introduced the notion of prototype: the prototype is the best representative of the category. However, what is prototypical for some may not be for others, so prototypicality differs across individuals, depending on their particular experience. In this view, membership in a category is no longer a yes/no question but depends on the object's degree of similarity to the prototype, with certain exemplars being central to the category, whereas others are more peripheral. As we shall see later, the notion of prototype proposed in cognitive psychology is similar to the notion of internal standard described by Kreiman et al.[3,4,13]

TRAINING-BASED REINFORCEMENT OF INTERNAL STANDARDS

Unstable and listener-specific prototypes

The lack of reliability in perceptual evaluations of dysphonia is largely dependent on what strategies and mechanisms are used by listeners to classify voices, particularly the auditory standards used by the judges. Kreiman et al[3] introduced the notion of internal standard, which is equivalent to the notion of prototype used in cognitive psychology: each listener judges the quality of a voice by comparing it to his or her internal auditory standard or prototype, which is based on what the person thinks a normal or dysphonic voice sounds like. It is the perceptually estimated distance between one's internal standard and the voice heard that determines the degree of severity assigned to the voice. As a whole, however, such internal standards are listener specific and are more or less precisely defined in accordance with the listener's perceptual experience with dysphonic voices.

External anchoring based on comparison

An alternative assessment method was used by Gerratt et al,[13] who proposed replacing internal standards by a set of external anchors or perceptual references. By supplying a voice scale that is constant and the same for everyone, this method allows listeners to categorize voice samples by comparing them to the reference set.[14] In one study, these authors demonstrated the effectiveness of using an externally anchored scale to evaluate voice roughness.[13] Their use of synthesized voices[15–18] seemed justified because it could provide a range of voice samples representing the diverse manifestations of dysphonia, both in terms of quality (breathiness and roughness) and severity. But a paradigm with synthesized and/or pseudonatural voice anchors (To obtain the full range of dysphonic voice samples, Chan and Yiu[17] had a healthy speaker simulate various degrees of roughness and breathiness, hence our use of the term "pseudonatural.") leaves much to be desired, not only because synthesized voices are too artificial to be compared with natural voices but also because difficulty calibrating stimuli derived from natural voices makes it hard to obtain a representative set of voice samples. Last, a method based on systematic comparisons with external references is far removed from everyday speech perception and therefore does not leave listeners in a position to judge for themselves after training.

Our proposal: internal anchoring via training on natural voices

The use of external anchors can generate "unnatural" situations for speech perception. In an attempt to avoid this problem, we designed a training method similar to the ones developed for learning new words or phonemes,[19,20] which do not supply external references to subjects. This approach is similar to the approach of Martin and Wolfe[15] and Chan and Yiu[18] but differs from theirs by the fact that our training protocol makes use of real dysphonic voices rather than synthesized or simulated ones.

For the inexperienced listeners, the experiment took place in three phases: (1) a pretest for measuring the listener's initial performance on dysphonia categorization, (2) a training phase, and (3) a posttest to detect any changes resulting from the training. To assess the learners' final performance, we compared them to a group of experts who were given a test identical to the posttest taken by the inexperienced participants. Two questions were raised in this study. The first was aimed at finding out whether inexperienced listeners can learn to categorize dysphonia severity. The second was aimed at

determining whether our training protocol could result in expert-level performance.

MATERIALS AND METHODS

Assessing categorization performance

We chose Grade, Roughness, Breathiness, Asthenia, Strain scale of Hirano[11] for the perceptual evaluation of dysphonia, the most widespread scale in use today. We decided to limit the assessment to dimension G, which measures overall dysphonia severity. This reduced the categorization process to four classes: normal voice (G0), slightly dysphonic voice (G1), moderately dysphonic voice (G2), and severely dysphonic voice (G3). One of the limitations of this choice is that dimension G does not account for the multidimensional aspects of vocal dysfunction. Indeed, a moderately dysphonic voice (G2) may also be breathy (eg, G2 R0 B2), rough (eg, G2 R2 B0), or both (eg, G2 R2 B2), and other dimensions may be at play, including bitonality, tenseness/laxness, and temporal instability. As a result, sound samples in the same class are likely to be heterogeneous from the acoustic standpoint. On the other hand, it has been shown that G is the most reliable dimension on this scale.[21]

Utterances

Our goal was to work in as natural a context as possible. We therefore eliminated sustained vowels, which are uncommon in natural speech and lead to the underestimation of dysphonia severity.[8] We also excluded spontaneous speech because the training protocol cannot be standardized. Our final choice was the reading aloud of Alphonse Daudet's "*La chèvre de Monsieur Seguin*" (Mr. Seguin's Goat), which has been used for about 20 years to record dysphonic patients at the Ear, Nose, and Throat Ward of the Timone University Hospital in Marseille.[22] Given that dysphonia can manifest itself temporarily and in a nonuniform way, we chose the sentence "*Il les perdait toutes de la même façon*" (He was losing all of them in the same way) because it contains a series of voiced/voiceless transitions and vocalic continuums, and its prosodic structure is such that the word "*toutes*" (all) is stressed. This structure is particularly interesting because the combination of an F0 rise (which is not always realized properly by dysphonics) and a (voiceless stop + vowel + voiceless stop + vowel) syllable structure is likely to reveal cases of laryngeal difficulty.

Speakers and corpus

Opting for natural samples as a methodological choice. A major difficulty in this type of experimentation is selecting samples that are representative of the categories to be used as the training material. In the study of dysphonia, the lack of a theoretical model poses a substantial problem when it comes to choosing speech samples. In fact—and this is the main problem faced today—there is nothing in the literature that allows us to define a normal voice, let alone a slightly, moderately, or severely dysphonic one. Martin and Wolfe[15] sidestepped this problem by synthesizing artificial voices while varying the jitter (instantaneous instability of the fundamental frequency) and the signal/noise ratio in the synthesized stimuli. Chan and Yiu[18] relied on the same principle by manipulating the amplitude of aspiration and the index of diplophonia on Klatt's HLSyn synthesizer. We took a different approach in our study to manipulate natural voices only.

Patients and speakers. We drew our data from the large number of dysphonic speakers and controls recorded by the Ear, Nose, and Throat Ward of the Timone University Hospital in Marseille or by the Neurology Ward of the University Hospital in Aix-en-Provence. The database includes samples from about 2500 dysphonic or dysarthric speakers.[22] For our categorization task, we confined the corpus to voices of women with various types of dysfunctional dysphonia resulting from nodules, polyps, cysts, or Reinke's edema. Speakers with no vocal disorders were added, giving us a selection of 400 female voices.

For each recording selected, we extracted the target sentence chosen for our training protocol ("*Il les perdait toutes de la même façon*"). Each vocal excerpt was then categorized by an automatic dysphonia analyzer. The principle of this analyzer consists in adapting a classical speaker recognition system to voice quality classification.[23] A speaker recognition system is a supervised classification system able to differentiate speech signals into classes. In our case, a class corresponds to either a grade of dysphonic patients or normal subjects. The speaker recognition technique used is based on a Gaussian Mixture Model (GMM)-based approach, which is the state-of-the-art for speaker recognition. This approach needs three phases: parameterization, model training, and classification. Parameterization consists in extracting information from speech signal. Each signal frame (20 milliseconds) is characterized by MEL-frequency cepstral coefficients that contain spectral information. The class model is learnt using data from a set of speakers who belong to the same grade. This training phase is based on the expectation maximization/maximum likelihood algorithm, able to extract statistical information for each class. Obviously, the voices used for the class training are not included in the test set to differentiate voice quality classification from speaker recognition. During the classification phase, an input signal is presented to the system, compared with the model of each class and assigned to the closest class in terms of similarity measure (likelihood). This device has proven its ability to correctly rate dysphonia severity in approximately 80% of the cases.[23] In parallel and independently, we asked a panel of three experts to do a blind perceptual evaluation of the same samples. These speech therapists and voice pathologists were accustomed to hearing dysphonic voices. Only those samples for which all three experts and the automatic device gave the same rating were retained. This procedure (described in detail in Ref. [24]) guaranteed that the voice samples used were representative and sufficiently diverse to reflect the full range of natural manifestations of dysphonia.

This gave us a complete corpus containing 142 voices, divided up as follows: 33 grade G0 voices, 32 grade G1 voices, 35 grade G2 voices, and 42 grade G3 voices. All voice samples used in the present experiment were extracted from this final corpus.

EXPERIMENT 1: CAN INEXPERIENCED LISTENERS LEARN TO CATEGORIZE DYSPHONIA BY SEVERITY?

Listeners

Thirty-eight listeners who had no experience in listening to dysphonic voices participated in the experiment. All were native speakers of French and reported having no hearing problems. For the purposes of the experiment, the participants were divided into four groups (three groups of 10 and one group of eight).

Procedure

The experiment was run under *PERCEVAL* software (LPL Dev., www.lpl-aix.fr) with its LANCELOT extension.[25] It is a program to design and run perceptual experiments on a computer. Auditory stimuli are played randomly by blocks. Listeners' answers are automatically saved.

The first experiment was divided in three phases: a pretest to assess the listener's initial performance in dysphonia categorization, a training phase, and a posttest to check for learning effects (changes in dysphonia categorization brought about by the training protocol).

Pretest. On the pretest, the listeners heard a series of voices and had to assign one of the four grades to each one (G0, G1, G2, and G3). No feedback about the expected response was given. The pretest consisted of a block of 20 voice samples, five per grade, presented in random order. To ensure that the mere fact of taking the pretest was not sufficient to improve performance via habituation to the various degrees of severity, the block was presented three times and we compared performance across repetitions.

Training. During the training phase, the listeners had to learn how to categorize new voices (ones not used on the pretest) by assigning one of the four grades to each sample heard. Insofar as grades G1 and G2 exhibit the greatest degree of variability, these grades were learned separately to reinforce the internal standard corresponding to each one. This procedure was inspired by a technique used in speech therapy in cases of auditory confusion between two phonemes: each phoneme is reinforced separately before it is opposed. The training phase was divided into four blocks (Table 1). Each block had six voice samples per grade, and the presentation order of the voices within each block was random. The first two blocks contained three grades each: G0, G1, and G3 for one block and G0, G2, and G3 for the other. The presentation order of the two blocks was counterbalanced across participants in such a way that G1 was always learned first for half of the participants and G2 was always learned first for the other half. The third block was made up of all four grades (G0, G1, G2, and G3) so as to oppose G1 and G2 within the same block. These first three blocks were presented twice in a row.

A trial ran as follows. Participants heard a voice and then saw the grade labels displayed on the screen. They had to click on the grade they assigned to the voice heard. Once they had responded, the expected grade appeared on the screen and the voice was repeated. At the end of each block, the participants were informed of their score on that block. On the last block (block 4), all four grades were presented again, but the participants were given no feedback about the expected responses and were not informed of their score.

Posttests. We used the same procedure on the posttest as on the pretest. This allowed us to check for learning effects, that is, any changes in our listeners' ability to categorize dysphonia severity. To determine whether the learning effects were long lasting, the posttest was administered twice, once immediately after the training phase (T0), and then again a week later (T + 7). Like the pretest, the posttest was composed of a block of 20 voice samples, five per grade, repeated three times. To avoid improvements in performance between the pre- and posttest due solely to particularities of the voices used in each test, the 20 voice samples used on the pretest for half of the participants were used on the posttest for the other half and vice versa.

Results

Pretest. The scores obtained on the pretest are shown in Table 2. Analysis of variance (ANOVA) was conducted with grade (G0, G1, G2, and G3) and presentation (first, second, and third) as variables.

Only the main effect of grade ($F(3,111) = 50.83$, $P < 0.0001$) was significant. Neither the main effect of presentation ($F(2,74) = 0.07$, $P > 0.20$) nor the interaction between grade and presentation ($F(6,222) = 0.31$, $P > 0.20$) was significant. The scores obtained on each grade did not evolve across repetitions, which suggests that simply making voice quality judgments without expected response feedback does not suffice to improve dysphonia classification performance.

To clarify the main effect of grade, we conducted a series of pairwise comparisons with a Bonferroni-corrected alpha level of .008. The poorest scores were obtained for intermediate grades G1 and G2, although G1 voices turned out to be easier to categorize than G2 voices. Grade G0 gave rise to the best

TABLE 1.
Illustration of the Training Protocol

Block	Grade and Number of Voice Sample	Block Repetition	Feedback
1	6G0, 6**G1 or G2***, 6G3	Yes	Yes
2	6G0, 6**G2 or G1***, 6G3	Yes	Yes
3	6G0, 6G1, 6G2, 6G3	Yes	Yes
4	6G0, 6G1, 6G2, 6G3	No	No

* G1 or G2, depending on the participant group.

TABLE 2.
Mean Percentage of Expected Responses on the Pretest, by Grade

	G0	G1	G2	G3
1st presentation	85	49	38	68
2nd presentation	85	49	38	69
3rd presentation	88	49	37	67
All repetitions	86	49	38	68

performance, with an average of 86% expected responses. G3 scores were not as good as G0 scores but were better than G1 and G2 scores. Pairwise comparison results are displayed in Table 3.

Training. The results obtained on the training phase are presented in Tables 4 and 5. Because we were interested in the impact of our training on perceptual evaluation of dysphonia, we did not discuss the training results further and we directly compared pre- and posttest results to detect any changes resulting from the training protocol. Note however that the most striking result of our training phase is that as soon as intermediate grades G1 and G2 were mixed together in the same block, performance on these two grades dropped back down. At the end of the training, G1 and G2 voices remained the most difficult to categorize.

Posttests. The posttest results are presented in Table 6 For each posttest (T0 and T + 7), ANOVA was conducted with grade (G0, G1, G2, G3, and G4) and training order (G1 first and G2 first) as variables.

Immediately after training (T0). Only the main effect of grade ($F(3,108) = 82.05$, $P < 0.0001$) was significant. Neither the main effect of training order ($F(1,36) = 0.00$, $P > 0.20$) nor the interaction between grade and training order ($F(3,108) = 0.00$, $P > 0.20$) was significant.

To clarify the main effect of grade, we conducted a series of pairwise comparisons with a Bonferroni-corrected alpha level of .008. As on the pretest, performance was significantly better for grades G0 and G3. Contrary to the pretest, posttest performance for grade G3 voices was significantly higher than for grade G0 voices, and G1 and G2 scores did not differ significantly. Pairwise comparison results are displayed in Table 3.

To compare the pretest and immediate posttest scores, an additional ANOVA was conducted with test (pretest and post-test) and grade (G0, G1, G2, and G3) as variables. The main effect of test was significant ($F(1,37) = 24.05$, $P < 0.0001$). Overall, performance was better in the posttest than in the pretest. The main effect of grade ($F(3,111) = 110.24$, $P < 0.0001$) and the interaction between grade and test ($F(3,111) = 14.81$, $P < 0.0001$) were significant. This interaction showed a clear-cut improvement between the pretest and posttest for grades G2 ($F(1,37) = 19.52$, $P < 0.0001$) and G3 ($F(1,37) = 46.38$, $P < 0.0001$).

A week after training (T + 7). One participant did not come to the posttest a week later, so scores were calculated for 37 listeners only. Note that this participant was also discarded from the following analyses comparing score on posttest a week later to scores obtained on pretest and posttest immediately after training. Only the main effect of grade ($F(3,105) = 77.04$, $P < 0.0001$) was significant. Neither the main effect of training order ($F(1,35) = 0.89$, $P > 0.20$) nor the interaction between grade and training order ($F(3,105) = 0.22$, $P > 0.20$) was significant.

To clarify the main effect of grade, we conducted a series of pairwise comparisons with a Bonferroni-corrected alpha level of .008. Once again, performance was significantly better for grades G0 and G3. Grade G1 and G2 voices remained the most difficult to categorize, and no significant difference was found between these two intermediate grades. Pairwise comparison results are displayed in Table 3.

To compare the pretest and posttest taken a week later, an additional ANOVA was conducted with test (pretest and post-test) and grade (G0, G1, G2, and G3) as variables. The main effect of test was significant ($F(1,36) = 13.78$, $P < 0.001$). Overall, performance was better in the posttest than in the pretest. The main effect of grade ($F(3,108) = 107.57$, $P < 0.0001$) and the interaction between grade and test ($F(3,108) = 09.59$, $P < 0.0001$) were significant. Again, this interaction showed a clear improvement between the pretest and posttest for grades G2 ($F(1,36) = 09.90$, $P < 0.001$) and G3 ($F(1,36) = 34.17$, $P < 0.0001$). Hence, the improvements noted for grades G2 and G3 were still present after a 7-day delay.

TABLE 3.
Pairwise Comparison Results Testing the Main Effect of Grade With a Bonferroni-Corrected Alpha Level of .008

	G0	G1	G2
Pretest			
G1	$t(37) = 8.16$, $P = 0.0000$		
G2	$t(37) = 10.02$, $P = 0.0000$	$t(37) = 3.06$, $P = 0.004$	
G3	$t(37) = 3.98$, $P = 0.0003$	$t(37) = 4.18$, $P = 0.0002$	$t(37) = 10.83$, $P = 0.0000$
Posttest (T0)			
G1	$t(37) = 8.83$, $P = 0.0000$		
G2	$t(37) = 6.97$, $P = 0.0000$	$t(37) = 1.55$, $P = 0.13$	
G3	$t(37) = 5.20$, $P = 0.0000$	$t(37) = 14.68$, $P = 0.0000$	$t(37) = 15.98$, $P = 0.0000$
Posttest (T + 7)			
G1	$t(36) = 9.27$, $P = 0.0000$		
G2	$t(36) = 9.07$, $P = 0.0000$	$t(36) = 1.70$, $P = 0.10$	
G3	$t(36) = 0.71$, $P = 0.48$	$t(36) = 10.75$, $P = 0.0000$	$t(36) = 14.48$, $P = 0.0000$
Experts			
G1	$t(13) = 5.88$, $P = 0.0001$		
G2	$t(13) = 5.19$, $P = 0.0002$	$t(13) = 0.22$, $P = 0.83$	
G3	$t(13) = 3.39$, $P = 0.005$	$t(13) = 2.65$, $P = 0.02$	$t(13) = 2.31$, $P = 0.04$

TABLE 4.
Mean Percentage of Expected Responses, by Grade, for the Group That Learned G1 First During Training

Block	G0	G1	G2	G3
1				
1st presentation	74	65		89
2nd presentation	81	77		93
2				
1st presentation	98		88	83
2nd presentation	98		88	83
3				
1st presentation	87	63	63	82
2nd presentation	90	58	61	93
4 (No feedback)	80	54	58	88

Finally, a last ANOVA comparing the scores obtained immediately after training and those obtained a week later showed a main effect of test ($F(1,36) = 07.24$, $P < 0.05$) and a main effect of grade ($F(3,108) = 109.04$, $P < 0.0001$). The main effect of test was because of slight but significant decrease in performance between the two posttests. The interaction between grade and test was also significant ($F(3,108) = 3.22$, $P < 0.05$). Subsequent comparisons based on this significant interaction indicated a slight but significant decrease in performance for G1 voices ($F(1,36) = 4.14$, $P < 0.05$) and G3 voices ($F(1,36) = 09.63$, $P < 0.05$). For grade G2, a decline was also observed, but it was only marginally significant ($F(1,36) = 2.98$, $P = 0.09$). The increase in performance for grade G0 voices did not reach significance ($F(1,36) = 2.57$, $P = 0.12$).

Discussion

The pretest results indicated that the inexperienced listeners performed well on categorization of G0 and G3 voices. The high scores obtained on grade G0 were clearly because of the fact that unspecialized listeners have a stable internal standard for judging normal voices. This is not surprising because all listeners begin hearing this type of voice at birth. The relatively good performance obtained for G3 can be explained in terms of the location of this grade at the opposite end of the scale, which makes these highly degraded voices relatively easy to categorize. Grades G1 and G2 proved to be the most difficult to judge. This is not surprising either because inexperienced listeners, who rarely come in contact with dysphonic voices, probably do not have internal standards for the intermediate grades.

TABLE 5.
Mean Percentage of Expected Responses, by Grade, for the Group That Learned G2 First During Training

Block	G0	G1	G2	G3
1				
1st presentation	81		75	93
2nd presentation	81		75	93
2				
1st presentation	91	81		78
2nd presentation	96	89		85
3				
1st presentation	83	65	59	87
2nd presentation	92	69	60	93
4 (No feedback)	92	59	60	89

TABLE 6.
Mean Percentage of Expected Responses on the Posttests (T0 = Immediately After Training; T + 7 = a Week After Training), by Grade

	G0	G1	G2	G3
G1 learned first (T0)	84	53	59	97
G2 learned first (T0)	84	53	59	96
Total (T0)	84	53	59	96
G1 learned first (T + 7)	88	47	52	88
G2 learned first (T + 7)	88	49	55	94
Total (T + 7)	88	48	53	90

Comparison of the pretest and posttest scores showed that our training protocol was in fact effective—performance improved. The improvement seems to be the result of reinforcement of an internal standard for severely deteriorated voices (G3). Because of the proximity of G2 voices, this grade was also categorized better, having benefited from the reinforcement of a neighboring internal standard. This line of reasoning could also explain the lack of a posttraining improvement in G1 voice categorization. Because the training phase did not improve G0 perception, its neighboring grade G1 was no easier to categorize on the posttest than on the pretest. For grades G2 and G3, the improvements resulting from the learning were still present a week later. However, dysphonic voice judgment performance declined between the two posttests, suggesting that to achieve long-lasting effects, the training protocol must be repeated over time.

EXPERIMENT 2: DID OUR TRAINING PROTOCOL RESULT IN EXPERT PERFORMANCE?

Method

Fourteen expert listeners (voice and speech therapists) participated individually in the experiment. They had all been in practice for more than 2 years in the field of voice pathology (from 3 to 30 years of practice). This group is completely independent from the three experts who participated to the selection of the stimuli. For the experts, the experiment consisted solely of a posttest that was identical to the one taken by the inexperienced listeners in Experiment 1.

Results

The expert listeners' results are presented in Table 7, Figures 1 and 2. ANOVA conducted with grade as variable revealed a significant main effect ($F(3,39) = 15.24$, $P < 0.0001$). A series of pairwise comparisons with a Bonferroni-corrected alpha level of .008 showed that the highest scores were obtained on grade G0. The performance on grades G1, G2, and G3 did

TABLE 7.
Mean Percentage of Expected Responses, by Grade, for Expert Listeners

	G0	G1	G2	G3
Test (experts)	91	47	49	66

not differ significantly. Pairwise comparison results are displayed in Table 3.

A second ANOVA comparing the experts' scores with those of the inexperienced listeners' pretraining scores (Figure 1) revealed only a main effect of grade ($F(3,150) = 49.69$, $P < 0.0001$). Neither a main effect of group ($F(1,50) = 0.70$, $P > 0.20$) nor an interaction between group and grade ($F(3,150) = 1.09$, $P > 0.20$) was observed.

Surprisingly, our analyses revealed that the experts' scores did not differ significantly from the inexperienced listeners' pretraining scores, irrespective of the grade of dysphonia. We conducted a last ANOVA to compare the performance of our experts with that of our inexperienced listeners immediately after training (posttest scores, Figure 2). The main effect of grade was significant ($F(3,150) = 61.65$, $P < 0.0001$). Crucially, both the main effect of group ($F(1,50) = 16.79$, $P < 0.001$) and the interaction between group and grade ($F(3,150) = 9.88$, $P < 0.0001$) were significant. After training, inexperienced listeners performed better than expert listeners did on severely dysphonic voices (G3, $F(1,50) = 71.39$, $P < 0.0001$). Also, inexperienced listeners exhibited a marginally significant tendency to perform better than expert listeners on moderately dysphonic voices (G2, $F(1,50) = 3.35$, $P = 0.07$).

Discussion

The results of Experiment 2 showed that before training, inexperienced subjects already exhibited a high level of performance similar to that of experts, no matter what grade of dysphonia was being rated. The lack of a difference between expert and inexperienced listeners has been observed in the past[6,26] and casts doubt on the notion of expertise in blind perceptual evaluation tasks. It neither, of course, questions the notion of medical expertise in any way nor the ability of experts to make accurate descriptions of voice disorders. On the other hand, it seems clear that all listeners possess the ability to distinguish between normal, moderately degraded, and highly degraded voices, and that this ability does not necessarily improve with professional experience.

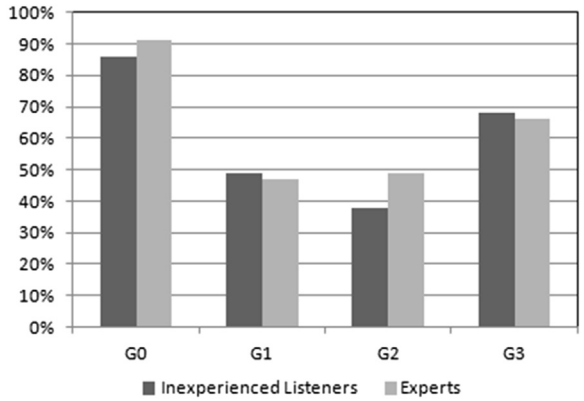

FIGURE 1. Comparison of scores of inexperienced pretraining listeners and expert listeners.

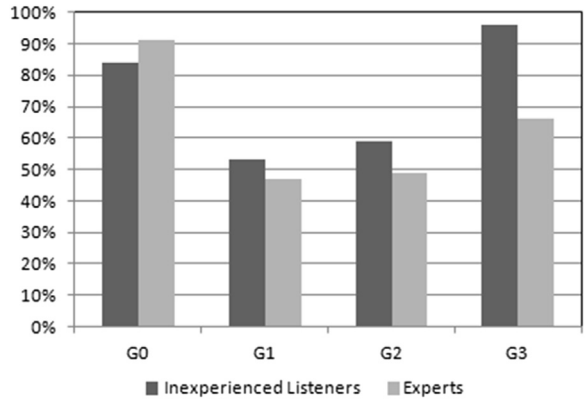

FIGURE 2. Comparison of scores of inexperienced posttraining listeners and expert listeners.

The inexperienced subjects performed better after training, and they even outperformed the experts for grades G2 and G3. We think that our training protocol allowed them to generate more robust prototypes, and especially, to build common references shared by all listeners. The fact that it was on extreme grade G3 that they improved the most, followed by grade G2 (by way of a proximity effect), prompts us to contend that above all, the training phase enabled them to calibrate the perceptual space, a calibration process that was lacking among the experts, who built their metric in an individual and perhaps idiosyncratic fashion.

CONCLUSION

Our training protocol—designed to teach listeners how to perceptually judge the severity of dysphonic voices—proved effective. One of its advantages is that it relies on natural speech samples, thereby facilitating the transfer of what was learned to subsequent assessments of patients' voices in a clinical setting. The voice samples chosen for each grade of dysphonia underwent a strict selection process based on multiple analyses and cross-analysis consistency. Only this type of sample selection method allows listeners to build a robust internal representation of the severity of dysphonia in its multiple forms.

The present study showed not only that it is possible for inexperienced listeners to learn to perceptually classify dysphonic voices by severity but also that they can do this so well that their performance equals or even surpasses that of expert voice listeners.

As a whole, listeners have a stable internal anchor for normal voices (G0), with correct identification rates between 80% and 90%. The two groups of listeners performed similarly, and their responses were stable, that is, experts and inexperienced listeners performed equally well and the training phase did not change the ratings for this voice grade. Severely dysphonic

voices (G3) were perceived moderately well by both experts and inexperienced pretraining listeners (66% and 68% correct, respectively). The scores of the latter group improved greatly with training, reaching a correct identification rate of more than 90%. Slightly (G1) or moderately (G2) dysphonic voices were not perceived very well neither by the experts nor by the inexperienced listeners before training (identification rates of less than 50%). Only the categorization of G2 voices benefited from the training phase, after which the mean score was as high as 60% for the inexperienced listeners, probably because of G2's proximity to G3, for which categorization scores improved considerably.

The difficulty observed on intermediate grades G1 and G2 suggests that categorizing dysphonic voices into four severity levels might not be an optimal approach. One can hypothesize that it might be more useful to reduce the perceptual scale to three categories (normal, moderately dysphonic, and severely dysphonic) because this corresponds better to the perceptual abilities of listeners.

The irregular improvements in performance after training—with the major effect observed for extreme grade G3—suggest that more than anything else, our training protocol enabled the learners to define a common internal standard. The absence of such a standard among experts can explain the unreliability of dysphonic voice judgments generally observed in the form of rating variability. It would be worthwhile to apply our protocol to experts, for whom it would act more as a calibration technique than a training device. Like Eadie et al,[26] we think that this type of protocol could be fruitfully used to train voice therapists.

Acknowledgments

We thank Maud Rouaze and Valérie Bokanowski for their contribution to this study. This work benefited from the support of the French government. It was carried out by Labex BLRI (ANR-11-LABX-0036), managed by the French National Research Agency (ANR), under the project title "Investments of the Future A*MIDEX" (ANR-11-IDEX-0001–02).

REFERENCES

1. Gerratt B, Kreiman J. Theoretical and methodological development in the study of pathological voice quality. *J Phon*. 2000;28:335–342.
2. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
3. Kreiman J, Gerratt BR, Precoda K, Berke GS. Individual differences in voice quality perception. *J Speech Hear Res*. 1992;35:512–520.
4. Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
5. Bele I. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573.
6. De Bodt MS, Wuyts FL, Van De Heyning PH, Croux C. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
7. Wuyts FL, De Bodt MS, Van De Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517.
8. Révis J, Giovanni A, Wuyts F, Triglia J. Comparison of different voice samples for perceptual analysis. *Folia Phoniatr Logop*. 1999;51:108–116.
9. Shrivastav R. Multidimensional scaling of breathy voice quality: individual differences in perception. *J Voice*. 2006;20:211–222.
10. Eadie TL, Baylor CR. The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *J Voice*. 2006;20:527–544.
11. Hirano M. *Clinical Examination of Voice*. Wien, New York: Springer-Verlag; 1981.
12. Rosch E. Natural categories. *Cognit Psychol*. 1973;4:328–350.
13. Gerratt BR, Kreiman J, Antonanzas-Barroso N, Berke GS. Comparing internal and external standards in voice quality judgments. *J Speech Hear Res*. 1993;36:14–20.
14. Fex S. Perceptual evaluation. *J Voice*. 1992;6:155–158.
15. Martin DP, Wolfe VI. Effects of perceptual training based upon synthesized voice signals. *Percept Mot Skills*. 1996;83(3 pt 2):1291–1298.
16. Gerratt BR, Kreiman J. Measuring vocal quality with speech synthesis. *J Acoust Soc Am*. 2001;110(5 pt 1):2560–2566.
17. Chan KMK, Yiu EM. The effect of anchors and training on the reliability of perceptual voice evaluation. *J Speech Lang Hear Res*. 2002;45:111–126.
18. Chan KMK, Yiu EM. A comparison of two perceptual voice evaluation training programs for naive listeners. *J Voice*. 2006;20:229–241.
19. Magnuson JS, Tanenhaus MK, Aslin RN, Dahan D. The time course of spoken word recognition and learning: studies with artificial lexicons. *J Exp Psychol Gen*. 2003;132:202–227.
20. Dufour S, Nguyen N, Frauenfelder UH. Does training on a phonemic contrast absent in the listener's dialect influence word recognition? *J Acoust Soc Am*. 2010;128:EL43–EL48.
21. Dejonckere PH, Obbens C, De Moor GM, Wieneke GH. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr (Basel)*. 1993;45:76–83.
22. Ghio A, Pouchoulin G, Teston B, et al. How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Comm*. 2012;54:664–679.
23. Fredouille C, Pouchoulin G, Ghio A, Révis J, Bonastre J, Giovanni A. Back-and-forth methodology for objective voice quality assessment: from/to expert knowledge to/from automatic classification of dysphonia. *EURASIP J Adv Sig Pr*. 2009;2009:1–14.
24. Ghio A, Dufour S, Rouaze M, Bokanowski V, Pouchoulin G, Révis J, Giovanni A. Perceptual assessment of dysphonia: a training protocol with natural speech. *Rev Laryngol Otol Rhinol (Bord)*. 2011;132:19–27.
25. André C., Ghio A., Cavé C., Teston B., PERCEVAL: a computer-driven system for experimentation on auditory and visual perception. Presented at: Proceedings of XVth ICPhS; 2003; Barcelona, Spain, 1421–1424.
26. Eadie TL, Kapsner M, Rosenzweig J, Waugh P, Hillel A, Merati A. The role of experience on judgments of dysphonia. *J Voice*. 2010;24:564–573.